

The Design of Virtual Audiences: Noticeable and Recognizable Behavioral Styles

Abstract

Expressive virtual audiences are used in scientific research, psychotherapy, and training. To create an expressive virtual audience, developers need to know how specific audience behaviors are associated with certain characteristics of an audience, such as attitude, and how well people can recognize these characteristics. To examine this, four studies were conducted on a virtual audience and its behavioral models: (I) a perception study of a virtual audience showed that people ($n = 24$) could perceive changes in some of the mood, personality, and attitude parameters of the virtual audience; (II) a design experiment whereby individuals ($n = 24$) constructed 23 different audience scenarios indicated that the understanding of audience styles was consistent across individuals, and the clustering of similar settings of the virtual audience parameters revealed five distinct generic audience styles; (III) a perception validation study of these five audience styles showed that people ($n = 100$) could differentiate between some of the styles, and the audience's attentiveness was the most dominating audience characteristic that people perceived; (IV) the examination of the behavioral model of the virtual audience identified several typical audience behaviors for each style. We anticipate that future developers can use these findings to create distinct virtual audiences with recognizable behaviors.

Keywords: virtual audience; expressive behavior; bodily expression recognition; simulated audience settings

1. Introduction

Virtual audiences can elicit responses in humans similar to those that are elicited by real human audiences (Slater, Pertaub, Barker, & Clark, 2006; Zambaka, Ulinski, Goolkasian, & Hodges, 2007). This is used in scientific research (e.g., Kelly, Matheson, Martinez, Merali, & Anisman, 2007), psychotherapy (e.g., Powers & Emmelkamp, 2008), and training (e.g., Bissonnette, Dubé, Provencher, & Moreno Sala, 2015), because virtual environments are easier to configure and control than the real world. While some applications aim for a neutral audience (e.g., Wallergard, Jonsson, Osterberg, Johansson, & Karlson, 2011), others may benefit more from an expressive audience. For example, the treatment manuals of exposure therapy (Heimberg & Becker, 2002; Hofmann & Otto, 2008) suggest controlling the audience attitude as an effective means of controlling anxiety in a public speaking scenario; studies on stress responses explored variations of stress tests using supportive and non-supportive audiences (Kelly et al., 2007; Taylor et al., 2010). As virtual audiences in public speaking scenario are becoming more widely used, e.g., as part of the Trier Social Stress Test (TSST) (Kirschbaum, Pirke, & Hellhammer, 1993), and in exposure therapy for social anxiety disorder, an empirically validated expressive virtual audience appropriate for these applications is needed.

When individuals are exposed to a virtual environment and perform in front of a group of virtual humans, their belief, anxiety, and performance can be affected. For example, Wallergard et al. (2011) suggested that virtual audiences as part of a stress test can indeed, like human audiences, induce stress. Aymerich-Franch, Kizilcec, and Bailenson (2014) used a virtual audience to study the effects of self-representation on public speaking anxiety. When presenting in front of a virtual audience, the individuals could see in a virtual mirror their virtual reflection which was manipulated to be similar or dissimilar to themselves. Others (Anderson et al., 2013; Hartanto et al., 2015; Morina, Brinkman, Hartanto, Kampmann, & Emmelkamp, 2015) focused on giving people the experience of performing in front of an audience as part of exposure therapy for individuals with social anxiety disorder. This experience has also benefited non-clinical applications. For example, Bautista and Boone (2015) let teachers be trained with virtual students to master their skills of content delivery and student management. Likewise, Bissonnette et al. (2015) trained performance arts students, in this case, young musicians to overcome their performance anxiety by performing in

front of a virtual audience. The information expressed by virtual audiences can be used for various purposes. For example, the virtual audience in a public speaking training system manifested different attitudes as feedback for the speech performance (Chollet, Sratou, & Shapiro, 2014). Supportive and non-supportive audiences have been used to evoke different levels of anxiety (Kelly et al., 2007; Taylor et al., 2010). Thus, the expressiveness of a virtual audience, i.e., what information a virtual audience can express and whether people can recognize the information, becomes a key question when designing virtual audiences.

As virtual audiences are made up of individual virtual humans, the first step in the development is the generation of individual virtual humans with believable behavior. Extensive work has been done in simulating such behavior. This work ranges from facial expression of emotion (Broekens, Qu, & Brinkman, 2012), head movement (Wang, Lee, & Marsella, 2011), to full body posture simulation (Chollet, Ochs, & Pelachaud, 2014; Xu, Pelachaud, & Marsella, 2014). Besides emotions, Chollet et al. (2014) and Hu, Walker, Neff, and Tree (2015) demonstrated that attitude and even personality of an individual virtual human can effectively be expressed by body language. To make the virtual characters believable, dynamic behaviors, i.e., displaying sequences of behaviors instead of still images, are often required. These sequences can be pre-scripted (Hu et al., 2015), computed by crafted rules that specify which behavior should be generated in a certain context based on psychological knowledge and literature (Bevacqua, Sevin, Hyniewska, & Pelachaud, 2012), or generated by statistical models that predict body postures based on observation (Chollet, Ochs, et al., 2014).

Besides the behavior as individual virtual humans, audience members also respond to each other's behavior. Although work has been done on crowd behavior (Thalmann & Musse, 2013) such as path planning and interaction between individuals of pedestrians, Kang et al. (2013) specifically had looked at the interaction behavior in an audience. According to their audience model, when an individual audience member is looking at an audience member in the neighborhood, the member in the neighborhood responds by looking back.

Among various public situations, public speaking is a common scenario occurring in everybody's life, e.g., delivering a business proposal, teaching in class, or giving a speech at a wedding. In public speaking situations, body language is a main channel of expression for audiences. Knowledge about this is therefore essential for developers to develop audiences that can be tailored for the need of users at run time. Currently, studies on the effects of virtual audiences often used three audience styles, described as positive, neutral and negative (e.g., Pertaub et al., 2002; Taylor et al., 2010). Their results showed the benefit and potential of varying audience styles. However, no explicit and unified descriptions or guidelines could be found for designing such virtual audiences. Therefore, it is still a challenge for future studies that needs either similar or different audience styles.

Limited research has been devoted to audience behavior in public speaking scenario. Poeschl and Doering (2012) and Tudor et al (2013) provided some guidelines for behavioral design of realistic virtual audiences. They observed the behavior of a typical audience in a lecture and explored the behavioral patterns such as frequency, duration, and postural sequence of certain behavior category, e.g. paying attention. Kang, Brinkman, Van Riemsdijk, and Neerinx (2013) proposed a parameterized audience model to generate expressive audience behavior for public speaking scenarios. The generated behavior was controlled by model parameters that defined the audience members' moods, attitudes, and personalities. They showed that the simulated audience using this model could behave expressively with regard to the audience attitude, and that the behavioral styles can be controlled by modifying the model parameters. Still, it is currently unclear about how an audience behaves underlying an audience style, e.g., a positive audience or a bored audience, and let alone which mood, attitude, or personality trait is associated with a specific audience style.

To simulate audiences for a variety of public speaking scenarios, more understanding about audience style and the relation with individual audience member characteristics is needed. These could be scenarios such as business people listening to an investment proposal pitch, employees assembled to hear the management announcement of potential layoffs, or students attending a Friday afternoon lecture who are eager to leave. Audiences in these settings clearly

behave differently. To simulate these audiences, a key question is how people differentiate between these audiences. Regardless of the narrative or the way people are dressed, are people able to recognize different audience styles in a similar way people are able to recognize different facial expression independent of the context, such as anger or sadness? And what are these audience styles?

To address these issues, the work presented in this paper uses an existing virtual audience environment (Kang et al., 2013) to address four questions: (1) what variations in audience characteristics, in particular, mood, personality, and attitude, result in perceivable variations in audience behavior? (2) What combination of individual audience members' characteristics do people use to design prescribed audience styles? (3) What audience styles do people recognize and (4) what are the typical audience postures and behaviors associated with specific audience styles? To answer these questions the paper first describes a paired comparison perception experiment, which is a classic psychophysical method that was used to determine peoples' sensitivity in noticing a specific quantitative difference in an audience characteristic, e.g., higher or lower arousal. After identifying which audience characteristic resulted in noticeable audience behavior differences, people were invited to use these characteristics to design audiences for a set of public speaking scenarios. Clustering the audience scenarios based on the similarity of the characteristic settings resulted in five audience styles. Videos of virtual audiences were made for each style, and people were invited to match audience style description to each video. The last step of the study was to examine the parameterized audience model and identify specific audience postures and behaviors that were characteristic for the behavioral styles.

2. Virtual Audience Model and Simulation

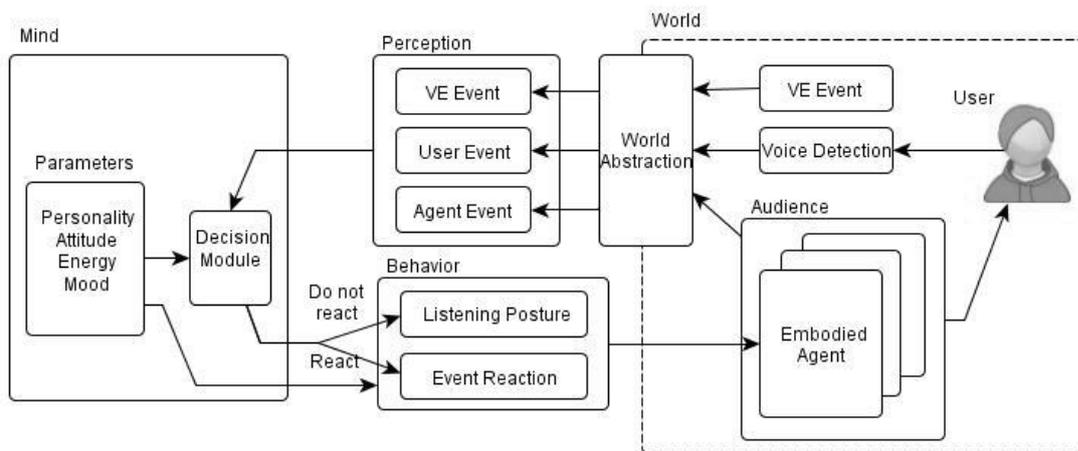


Fig. 1. Framework of the virtual audience simulator

The work in this paper revolves around a parameterized audience model (Fig. 1) (Kang et al., 2013) that underlies an audience of virtual humans in a virtual environment. This is a probabilistic model abstracted from observation of real human audiences. Behaviors of real audiences were recorded when they were listening to presentations on a topic they were interested in, were critical about, found boring, and were neutral about. The audience corpus (Kang, 2013) consists of 9600 coding units with a sampling interval of two seconds, specifying head, gaze, arm, hand, torso, and leg positions. To obtain a parameterized model, additional data was also collected about the audience members' personality (extroversion, agreeableness, openness, neuroticism, and conscientiousness), attitude towards the topic (interest, approval, eagerness for information, criticism, and impatience), mood (valence, arousal, and dominance), and energy level. To generate audience behaviors, 59 unique postures representing 80% of the corpus were grouped into 15 posture categories based on the probabilities of postures that would follow the current one in the observation corpus. The parameter data was used to train a series of logistic regression prediction functions for these 15 posture categories. Once the category is determined, the final posture of the virtual human is determined by random selection according to transition matrix of postures of the specific posture group. The full body posture of the virtual audience

was updated every two seconds. The audience model also includes event response, for example, turning head if a phone goes off, or looking back when another audience member is looking at the virtual human. Full details about the virtual audience and the parameterized audience model can be found in the work by Kang et al. (2013). Setting these parameters creates different audience styles. Kang et al (2013) showed that people could recognize different audience attitudes and perceive different degrees of attitudes in the audience simulation. In an attempt to make the mood of a virtual human more recognizable, the audience model in this paper was extended with facial emotion expression using the facial expression tool (Broekens, Qu, & Brinkman, 2012), which was directly controlled by the mood status of the virtual human. The facial expressions were static unless the mood status changed.

3. Study I: Perception of Changes in Parameters

3.1 Research questions and design

Kang et al (2013) report that when people were asked to describe freely various audiences or rate their characteristics such as attitude, personality, and mood, no expected difference were found for manipulations of parameters such as extraversion, valence, or arousal. Although changes in these parameters led to observable changes of audience behavior, it was not clear whether people were actually able to recognize the changes in parameters. Thus, this study addresses the question what audience characteristics people can perceive. To answer this question, pairwise comparison, a classic psychophysical method, was therefore applied. This method provides more precise results in interval scales than a direct scaling, because it transforms the scaling task, which is difficult for humans, into a comparison task (Engeldrum, 2000; Rajae-Joordens & Engel, 2005). For example, instead of being asked to rate directly the intensity of a specific characteristic of a virtual audience, participants are presented with two virtual audiences at the same time and asked which one they perceive to have a higher intensity of the characteristic.

3.2 Experiment settings

To determine perceivable audience characteristics, the first study investigated people's perception of the individual parameter changes. As some behaviors might only emerge when multiple parameters were changed, due to the correlation between the parameters in the observation corpus, the study also investigated people's perception when multiple parameters changed together as a group.

The parameter groups were determined by a principal component analysis on all the parameter data collected in the study by Kang et al (2013). The details of the analysis and grouping of the parameters are explained in Appendix A. The parameters with similar factor loadings, which indicated high correlations with each other, were grouped (Table 1), namely three independent parameter groups (IG1, IG2, and IG3), two independent single-parameter groups (IP1 and IP2), and three dependent parameters (DP1, DP2, and DP3). The independent groups correlated with different single factors, and the dependent parameters correlated with multiple factors. Therefore, the value of each parameter in the experiment was set by its correlated factors, i.e., when the factors were set, the values of the parameters were set.

The study investigated people's perceptions of the effects of 13 parameters in the model (Table 1) in two conditions: individual parameter adjustment and grouped parameter adjustment. Taking the perception of variation in the Interest parameter as an example here, in the individual parameter condition, only the Interest parameter was modulated, while in the grouped parameter condition three parameters of IG2 (i.e. impatience, eagerness for info, and interest) were modulated together. As the single-parameter groups (IP1 and IP2) only contained one parameter respectively, they were only included as individual parameters in the experiment. Additionally, in the grouped parameter condition, a combined question was used for independent parameter groups to see whether some parameters can be reduced to one control. For example, instead of only considering Interest, the participants were asked to give their overall opinion on the audience's Patience, Eagerness for information, and Interest together. For dependent parameters like Approval, no additional questions were asked. In both individual and group adjustment conditions, participants were asked to compare a few pairs of simulations in which the corresponding parameter or parameter group was set at

different levels. Because participants' task load can be extremely high when the number of stimuli is large, the number of stimuli should be set as small as possible. To exclude unnecessary levels, the number of levels for each parameter or parameter group was determined in a pilot study. In the pilot study, the first author, who was regarded as the expert of this model, attempted 12 times to differentiate between pairs of simulations set at ten different levels, the maximum supported by the model. Accordingly, the maximum number of levels that the first author was able to recognize significantly ($p < 0.05$) was employed in the real perception experiment (Table 1). For the combined question in group adjustment condition, the number of levels was determined by the parameter in the group with the fewest number of levels, e.g., three levels for the Interest group (IG2).

Table 1 Grouping of audience parameters and experiment settings

Parameters	Correlated factors	Grouping result	Number of levels for experiment	
			individual	group
Extraversion	1		3	3
Agreeableness	1	IG1	3	3
Conscientiousness	1		3	3
Openness	1, 4	DP1	3	3
Impatience	2		3	3
Eagerness for info	2	IG2	3	3
Interest	2		5	4
Approval	2, 3, 4	DP2	3	3
Dominance	3	IG3	3	3
Valence	3		7	4
Neuroticism	1, 3	DP3	3	3
Criticism	4	IP1	3	*
Arousal	5	IP2	5	*

Note: the loadings of the correlated factors are larger than 0.4.

* This parameter was not manipulated in a separate group condition.

3.3 Material and measures

An audience simulation was developed using the audience model mentioned in Section 2. To express mood better, facial expressions were added to this simulation using the facial expression tool (Broekens et al., unpublished results). Thus, the mood states of the audience affected not only the bodily responses, but also the facial expressions. To control the software load, only the audience members in the first row showed facial expressions in the audience simulation. An executable program was made for the pairwise comparisons, displaying two audience simulations side by side. The perception evaluation of parameters and parameter groups was conducted one by one separately, and the order was randomized. To evaluate the perceptions of one parameter (group) in N levels, N audience simulations were prepared with the parameter (group) varying from the lowest level (0) to the highest level (10) in the model. The program displayed sequentially one of all possible pairs (i.e., $N(N-1)/2$) of N audience simulations. The order of those pairs and the position (i.e. left or right) of the two audiences in each pair were randomly generated. For the evaluation of each parameter (group), a corresponding question was always displayed on the top, in the following form:

Which audience is more X?

According to the examined parameter (group), X was an adjective or a phrase, from the following list: pleased, aroused, dominant, open, conscientious, extroverted, agreeable, emotionally stable, positive towards the speech, interested in the talk, eager to get information, critical, and patient. All these terms corresponded to the audience parameters and were explained in an additional paper explanation card that included the definitions of the three dimensions of mood (Albert Mehrabian, 1996), the Big Five Factor of personality (Pervin, Cervone, & John, 2005), and the audience attitude questionnaire (MA) used in the previous study (Kang et al., 2013). For the combined question in the group condition, X was a combination of several terms which were involved in one group, e.g., “patient, interested, or eager to get information” for the Interest group.

3.4 Procedure

Hall (1984) suggested females were better decoders of nonverbal behavior. Thus, the gender of participants was balanced so that gender difference in perception of nonverbal behavior could be examined. Twenty-four participants (12 females and 12 males) were recruited throughout the university campus to evaluate the audience model. Their age ranged from 24 to 41 years with a mean of 28.5 (SD =3.4) years. Each participant was asked to watch pairs of the audience simulations displayed on two desktop displays (iiyama ProLite E4315) respectively. Each pair of audience conditions displayed simultaneously for 20 seconds. After the simulations stopped, the participant was asked to answer the question which audience was more X, and then the next pair was displayed. When finishing all the comparisons for an individual parameter or a parameter group, the participant was asked about the rationale for the choices in the comparisons.

3.5 Analysis and Results

The pairwise comparison data was analyzed using the method described in the study by Rajae-Joordens and Engel, (2005) (also described in Appendix B). The analysis was based on the Thurstone model (Thurstone, 1927), which provides scales on the differences people perceive among the stimuli. As post hoc analyses, multiple comparisons between simulation pairs were also conducted to examine whether these differences were significant. Confidence intervals of the differences, corrected by Scheffe’s method, were used as the criterion for significance. Only when there were significant differences between the levels of a parameter, the adjustment of the parameter (group) was regarded as perceivable. Fig. 2 shows the perceptual scales of all parameters and parameter groups. The scale was transformed that the value for the lowest level is always 0.0. These perceptual scale values show the relative locations of different levels of a parameter (group) on a psychological scale. That is, a value can be interpreted in terms of deviations from the values of other stimuli, and the deviations or differences follow a standard normal distribution. Taking Level 7 of valence in the individual condition for example (Fig. 2a), it was about one standard deviation from Level 4 and about one and a half standard deviations from Level 2. The parameter levels were grouped by ovals according to the confidence interval test. Thus, the levels were perceived significantly different if they belong to the non-overlapping area of two different ovals on one scale. If some ovals are overlapping, the levels belonging to the overlapping part (e.g. Level 2 and 3 in Fig. 2a, individual condition) do not show significant difference.

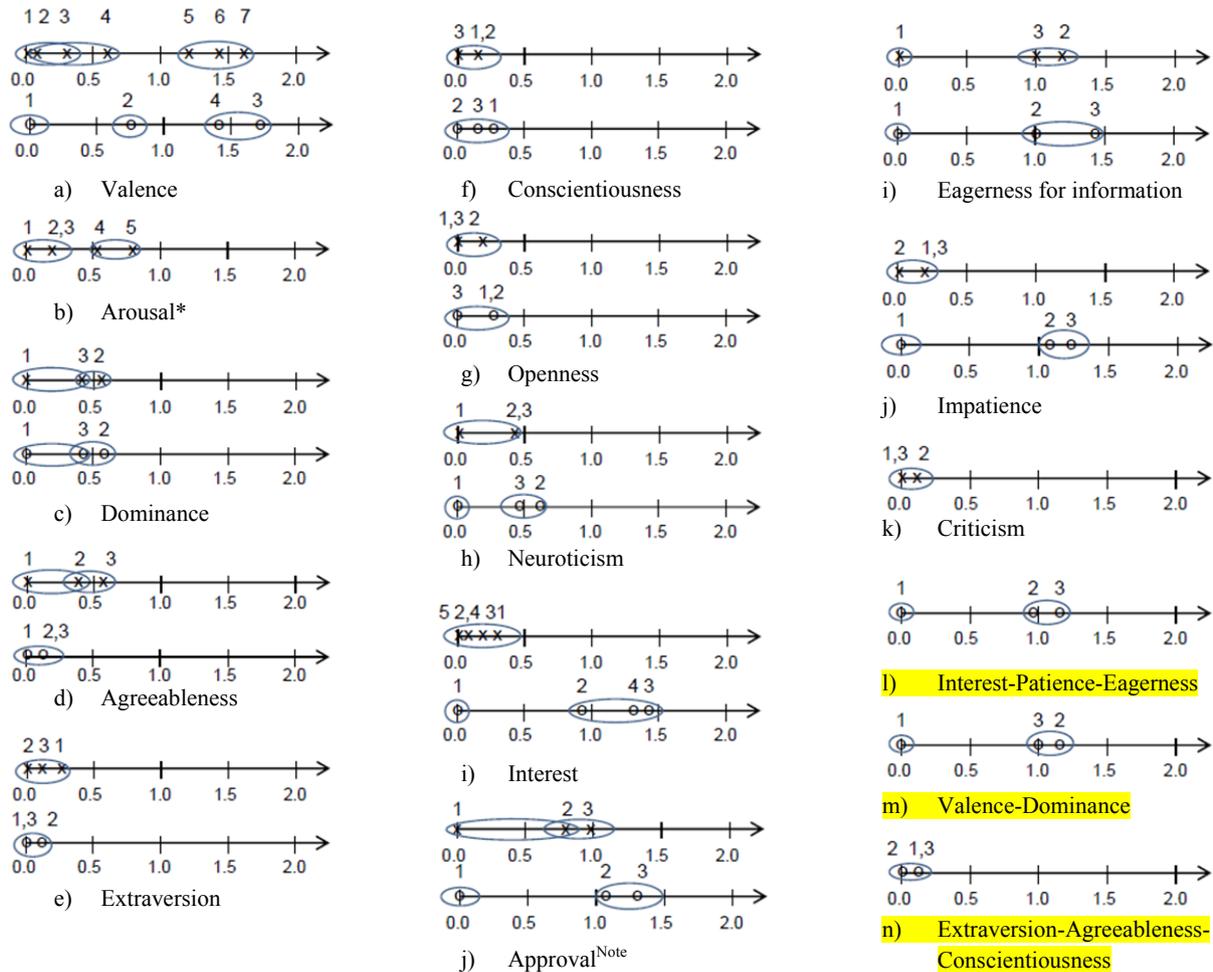


Fig. 2. Perceptual scales of different levels of all parameters and parameter groups. The crosses (×) on the scales are the results from the individual adjustment conditions; the circles (○) on the scales are the results from the group adjustment conditions. The levels are annotated by the numbers above the scales. The ovals demonstrate the statistical differences between the levels. Note: The figure for the group adjustment condition only includes the data of females instead of the whole sample (males and females).

The results showed that people could differentiate between two levels for the perceivable parameters and parameter groups, with an exception of three levels for the perception of Valence. The mood states were well recognized in both individual adjustment and group adjustment conditions. For personality dimensions, agreeableness was recognized in individual adjustment condition, and neuroticism was recognized when adjusting the corresponding parameter group. The attitude items (i.e., Interest, Approval, Eagerness for information, and Impatience) were mostly perceivable in the group adjustment condition, while, in the individual adjustment condition, only Eagerness for information was perceivable. Thus, the audience behavior was more expressive in the group adjustment condition than in the individual adjustment condition for the attitude items.

To examine whether gender affects the perception, the Thurstone model of the perception for each parameter or parameter group was extended into a generalized linear model (GLM), taking both the perceptual scales and gender as the independent variables. The model fit was compared between the GLM and the Thurstone model (as described in the study by Rajae-Joordens and Engel (2005)) respectively for each parameter or parameter group. Although males did not perceive the different levels of Approval as the females did in the group adjustment condition (Fig. 2j),

no significant difference (p values ranging from 0.08 to 0.99, with a mean of 0.52, $SD = 0.29$) was found between the results of males and those of females for any parameter (group).

4. Study II: Expressive Audience Design

Expressive virtual audiences are needed in various scenarios for different applications. Descriptions of audience behavior in these scenarios may provide direct information for the behavioral design. To this end, a design experiment was conducted to collect people's opinion on how audiences behave in different situations.

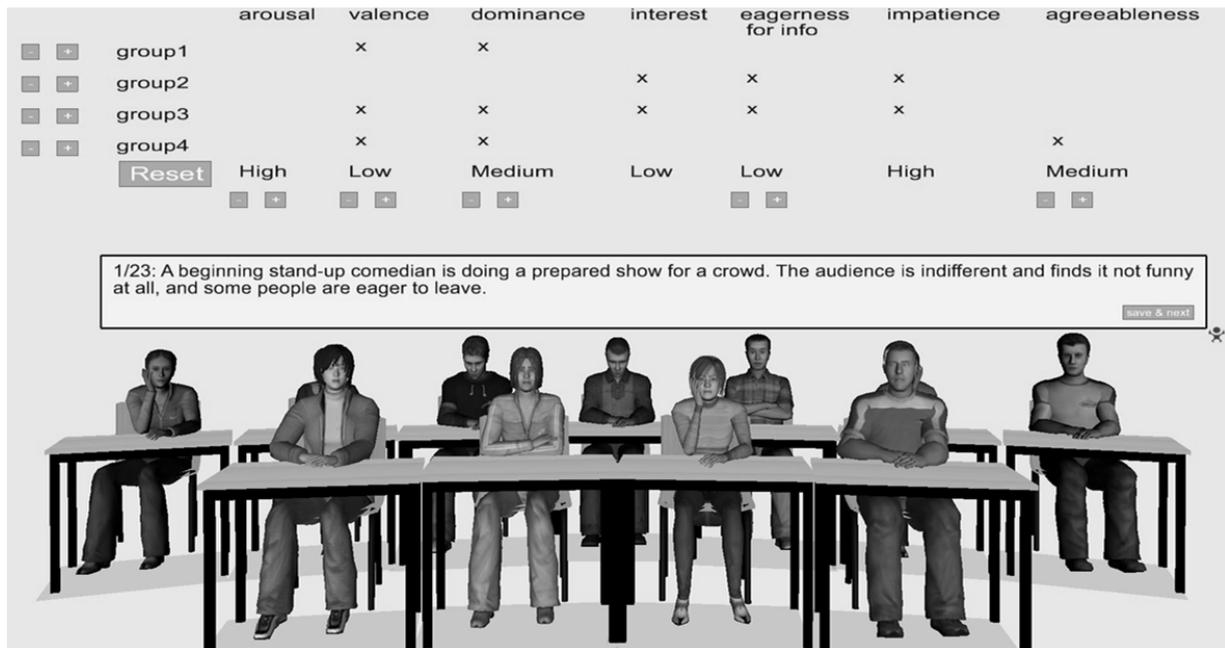


Fig. 3. A screen shot of the design interface.

4.1 Material

A design interface (Fig. 3) with one audience simulation and parameter controls was prepared so that participants could see the behavioral change when manipulating the audience parameters. The audience simulation model was similar to the one used in Study I. Unlike the study I where the virtual audience sat in a classroom, the audience for study II were seated in two rows in an arc with a blank background rendered in light blue. Without a specific background setting, the audience could fit into more scenarios. Controls of parameters were provided according to the results of the previous perception experiment (study I). Only the controls of parameters that were recognized in the previous study were provided so that the effect of parameter adjustment on the audience behavior was noticeable. In addition, participants could control parameters individually or as a group (i.e., some parameters increase or decrease together), according to the conditions in which the parameters were recognized. Thus, there were five controls for parameters which were recognized individually, namely, valence, arousal, dominance, eagerness for information, and agreeableness, and four controls for perceivable parameter groups, corresponded with Interest group (IG2), Valence-Dominance (IG3), Approval (DP2), and Neuroticism (DP3).

All the parameter controls allowed 3-level adjustment, i.e., low, medium, and high, and 0, 5, and 10 as the parameter value in the model, so that participant could always set medium level for the parameters. Although the perception experiment showed that people might only perceive two levels of some parameters, 3-level controls could avoid forcing participants to choose either high level or low level, thereby avoiding biased results. A reset button was also provided so that participants can reset all parameters to the medium level.

Twenty-three different audience scenarios (see appendix C) were described for participants to design the audience behaviors. Here are two examples of the scenario descriptions.

- 1) *During a weekly school assembly for high school students, the administrator is talking about the new rules the students should obey. The students find the rules much stricter than before.*
- 2) *At a booth of an exhibition, an exhibitor is introducing a new product. People follow the explanation and find the design innovative.*

4.2 Procedure

24 participants (12 females and 12 males) from 24 to 33 years old ($M = 27.5$, $SD = 2.5$) were recruited in the university. They were from seven different countries (14 Asians and 10 Europeans). They were asked to manipulate the audience behavioral styles by regulating some audience parameters so that the behavior matched the scenario descriptions. The paper explanation card about audience parameters was also given to the participants. The stimulated audience and the manipulation interface were shown on a TV (LG 42lm3450) about one meter from the position of participants. Before the experiment, participants practiced on manipulating the parameters to get some idea on how they could influence the behavioral styles. The order of scenario descriptions was randomized and participants were not allowed to go back to previous description and change the settings. After saving a setting, a new scenario description was given and all parameters were reset to the medium level.

4.3 Analysis and Results

The first step of the analysis was to examine whether there was similarity in the way participants had designed the audience behavior for each scenario. To check this, the consistency of the settings for each scenario across the 24 participants was investigated. If the design of one scenario was inconsistent, it was expected that the adjustment options (i.e., high, medium, and low levels) for each parameter in this scenario were random, i.e., equally selected by the participants. Hence, the Chi-square tests of goodness-of-fit were performed for each parameter against an equal distribution of the three options to determine whether the three options were equally preferred. Besides the five individually controlled parameters mentioned in Section 3.1, the settings of Interest were also examined because Interest is affected by both Interest group and Approval so that its value can be different from that of Eagerness for information. The test results (Table 2) showed that most settings (122 out of 132) were significantly different ($p < 0.05$) from a random setting, and each scenario setting had at least three out of six parameters that were consistent across the participants.

The next step was to analyze how these obtained audience settings varied between scenarios and how these scenarios could be clustered. To do this, the similarity between each pair of scenarios was investigated. The setting of each scenario consisted of six parameters and therefore could be considered as a point in a six-dimensional space. The setting similarity between a pair of scenarios was calculated by taking the Euclidean distance between the two points representing the scenario settings in this six-dimensional space. That is, the shorter the distance was, the more similar the two settings were considered to each other. To examine whether or not the two settings were similar, the observed distance was compared with the expected distance, i.e., the average distance between two random points. See Appendix D on how the observed distance and the expected distance were calculated.

Table 2 Results of chi-square tests on each parameter for each scenario, $\chi^2(2, N = 24)$

Scenario No.	χ^2 values for parameters					
	valence	arousal	dominance	interest	eager	agree
1	21.00	9.25	9.25	18.25	48.00	9.75
2	9.75	13.00	6.25	12.25	24.25	16.75
3	15.75	9.75	5.25*	7.00	3.25*	16.00
4	42.25	16.00	18.75	13.00	21.00	23.25
5	7.00	12.25	5.25*	10.75	24.25	14.25
6	14.25	10.75	1.75*	16.00	48.00	10.75
7	18.25	4.75*	6.75	13.00	32.25	12.25
8	18.25	19.75	18.75	18.25	13.00	9.00
9	12.25	24.25	24.25	14.25	48.00	12.25
10	42.25	28.00	12.25	12.00	16.00	15.75
11	21.00	19.75	0.75*	14.25	32.25	10.75
12	23.25	4.75*	5.25*	6.25	4.00*	12.25
13	14.25	12.25	3.25*	21.00	42.25	10.75
14	32.25	21.00	19.75	18.25	32.25	14.25
15	13.00	16.00	7.75	16.75	31.75	16.00
16	16.75	4.75*	1.00*	13.00	9.00	24.25
17	31.75	7.00	3.25*	12.00	21.00	14.25
18	21.00	12.25	21.00	18.25	37.00	7.75
19	19.00	4.00*	9.25	6.25	4.75*	5.25*
20	19.00	12.25	16.75	12.00	10.75	9.75
21	27.00	31.75	7.75	19.75	13.00	7.00
22	21.00	14.25	13.00	14.25	21.00	7.00
23	19.75	15.75	12.25	37.00	19.75	27.25

* The Chi-square value is less than 5.99, which is the critical value of $\chi^2(2)$ for $p = 0.05$. Thus, a result less than 5.99 indicates that the distribution was not found to deviate significantly from a random distribution.

As each scenario was designed by 24 participants, there were 24 observed distances to examine the similarity between the setting of one scenario and the setting of another scenario. Thus, these 24 distances were compared with the expected distance between a pair of scenarios by a one-sample t -test. If the observed distances between a pair of scenarios were significantly ($p < 0.05$) shorter than the expected distance, the two scenarios were regarded as similar to each other. According to the results, the scenarios were grouped into one category when they were similar to each other by showing significantly shorter distances between each other than the expected distance. The grouping results (Table 3) show that the scenario settings were grouped into five categories. The audience settings for scenarios 3 (the impractical business proposal) and 12 (the funeral eulogy) were excluded from the grouping because they did not show similarity in terms of short distance with any other scenarios.

Table 3 Grouping of the 21* audience scenario settings

Category	Scenario
A	1 (Promising business proposal), 4 (Best man’s talk), 6 (Tuesday morning lecture about exam), 7 (Attractive Tuesday morning lecture), 9 (Related Monday morning meeting), 10 (Funny show), 14 (Positive corporate report), 18 (Innovative design), 20 (A qualified interviewee)
B	2 (Potential business proposal), 21 (A not very satisfactory interviewee), 22 (Training plan), 23 (Hobby talk)
C	16 (Announcement of stricter rules), 17 (Budget cut)
D	11 (Not funny show), 13 (Souvenir introduction), 15 (Repeated rule announcement)
E	5 (Friday afternoon lecture), 8 (Unrelated Monday morning meeting), 19 (Looking around in an exhibition)

* Scenario 3 (Impractical business proposal) and 12 (Funeral eulogy) were excluded.

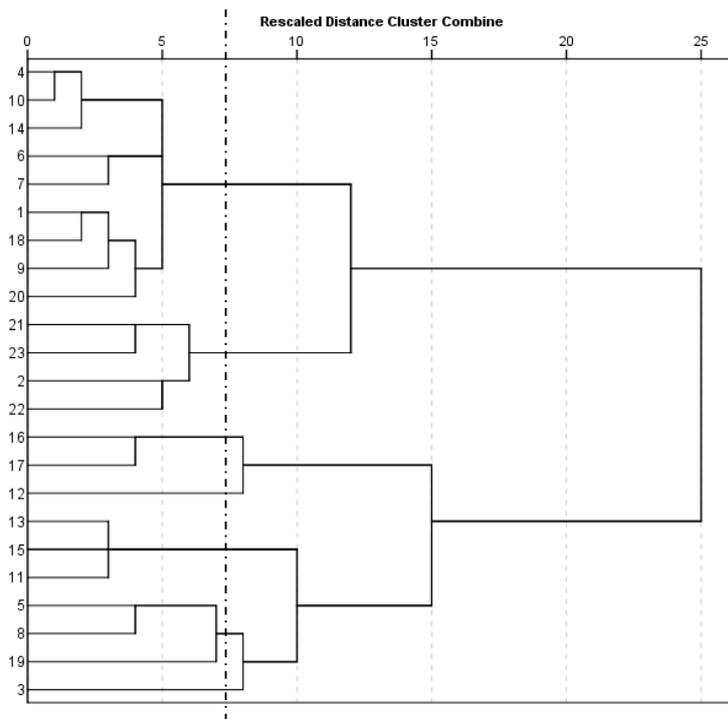


Fig. 4. The dendrogram using complete linkage. The dash line specifies the threshold distance of 7.5.

The categorization was also inspected by applying a clustering method. An agglomerative hierarchical clustering method with complete linkage was employed to group the audience settings using Euclidean distances between all setting pairs as the measure. The idea was to build a tree of data that successively merges similar groups of settings. The similarity between each setting pair was measured using Euclidean distances between the two settings. According to the dendrogram obtained in the hierarchical clustering process (Fig. 4), five categories were yielded by setting a distance threshold of 7.5 on a scale from 0 to 25, and Scenario 3 and 12 were excluded from the five categories. This categorization, if excluding Scenario 3 and 12, was exactly the same as that obtained by comparing the distances. Therefore, to obtain consistent setting features for each category, the audience settings were grouped into five categories, and Scenario 3 and 12 were excluded from the categorization.

The median settings of the five audience categories are shown in Table 4. The results provide an overview of the characteristics of each category. The difference in settings between the scenarios in different categories and similarity within a category suggest the potential existence of five distinct generic audience behavior styles.

Table 4 Median settings of the five audience categories

Parameter	Audience category				
	A	B	C	D	E
Valence	H	M	L	L	M
Arousal	H	M	H	L	L
Dominance	M	M	M	M	M
Interest	H	M	M	L	M
Eagerness	H	H	H	L	L
Agreeableness	H	M	L	L	M

L=0;M=5;H=10

5. Study III: Perception Validation of Audience Settings

The next step was to examine whether people were also able to recognize these five audience behavior categories. Thus, like other design-perception studies, (e.g., Xu, Broekens, & Hindriks, 2013), the design results need to be validated to ensure that people can recognize the designs. To validate the audience's behavior for different scenario descriptions, a perception experiment was conducted.

5.1 Material

The evaluation included the five settings (i.e., A, B, C, D, and E) found in the study II and a neutral setting as a baseline. To generate the different audience stimuli, the audience simulations from Study II were used. The five audiences were generated using the median settings (Table 4), and the neutral audience by setting all parameters to M. Three video clips of 30 seconds each were made for each audience setting from the audience simulation to avoid biased results, and two videos were randomly selected from the three and shown to participants. Thus, the evaluation consisted of 12 video clips: six settings and two clips for each. The 12 clips were displayed to participants in a random order to avoid the potential order effects.

5.2 Measures

A questionnaire was used to evaluate how well the virtual audience's behavior matched certain descriptions. The scenario descriptions were the same as those in the design experiment. The questionnaire was formulated as follows:

Which situations describe the audience in the movie best? Type A, B, C, D, or E? (Only one answer is possible.)

Type A: you may find such audiences in the following situations:

1) A person wants to start his own company and needs a sizable amount of investment money for this. He has an opportunity to introduce the investment proposal within 10 minutes to a number of business people, as they will consider whether or not they might invest in this new business opportunity. While listening, the investors find the proposal very promising.

2) *The best man is talking about some interesting story about the new couple at a wedding party. The people in the party are mostly the new couples' family members and friends. They are friendly and enjoy very much the stories.*

3) ...

Type B, you may find such audiences in the following situations:

...

The audience situations for each type were explained by listing the full descriptions of scenarios (Appendix B) that were categorized as that type (Table 3).

As the description for each scenario type is very long, short labels might be more convenient for future use. Thus, several words and short phrases were collected to label the audience types from nine people (six females and three males, ranging from 25 to 36 years old, $M = 29.6$, $SD = 3.6$). Two words were selected to label each type, shown in Table 5.

Table 5 Labels of different audience types

Audience type	Words and phrases collected	Label
A	² Attentive, paying attention, ² happy, ³ interested, ¹ related to the audience, ³ enthusiastic, ¹ engaged, ¹ open to ideas, ¹ easy, ¹ positive, ¹ intrigued, ¹ compliant, ¹ pleased	Interested and enthusiastic
B	¹ Hesitant, ² slightly negative, ² concerned, ¹ topic is related to audience, ¹ uncertain, ² cautious, ² mixed opinion, ⁴ critical, ¹ unclear	Critical and concerned
C	² Anxious, ¹ rebellious, ¹ fearful, ¹ not happy, ¹ worried, ¹ angry, ¹ betrayed, ² negative, ¹ opposite, ¹ tough, ¹ threatened	Anxious and threatened
D	¹ Inattentive, ¹ indifferent, ⁴ bored, ² impatient, ¹ annoyed, ¹ fed-up, ¹ disengaged, ¹ restless	Bored and impatient
E	¹ Distracted, ¹ impatient, ¹ not really interested, ² bored, ² uninterested, ¹ disengaged, ³ indifferent, ¹ restless	Indifferent and uninterested

Note: the superscript number ahead of a word indicates that the number of people who mentioned the word, e.g., for type A, “interested” was mentioned by three people.

To validate the labels, the participants were also asked to label the audience types by selecting one answer to the question as follows:

Which audience label describes the audience in the movie best? (Only one answer possible)

To check whether the participants could read English and take the survey seriously, the following open question was also included about the participants’ rationale for their choices.

Please fill in your rationale for your choices above in English. Please provide specific answers, instead of a generic answer for all videos.

5.3 Procedure

This survey was conducted online through Amazon Mechanical Turk among 101 people (51 females, 50 males) from the United States. Their age ranged from 18 to 64 years old ($M = 34.9$, $SD = 10.9$). The survey consisted of 12 parts. Each part contained one video clip and three questions; the participants were asked to watch the video and then

answer the description and label questions, as well as an open question about their rationale. The video could be replayed. Once finishing one part and continue with the next part, participants were not allowed to go back to the previous part. At the start, participants were informed that they would only receive their one-dollar reimbursement if they follow all the instructions.

5.4 Results

The result from one person was removed from the analysis because the answers to the open question were often obviously inconsistent with the choices of description and label. Before investigating how people perceived the videos, the descriptions and labels were examined to ensure that people could distinguish between the situation descriptions and between the audience labels. As the descriptions and the labels were expected to show one-to-one matches as designed, Cohen's kappa was calculated between the situation descriptions and audience labels. An agreement coefficient of 0.48 showed an overall significant ($p < 0.001$) association between the labels and descriptions. Fig. 5a shows the overview of the relationship between descriptions and labels. For descriptions A, B, C, and D, the most chosen label for each description was always the corresponding one, i.e., A, B, C, and D respectively. However, results did not always show one-to-one matches, e.g., descriptions D and E respectively matched with both labels D and E. Therefore, the descriptions and labels generally showed one-to-one matches, but the participants might not always be able to differentiate between the descriptions or labels of some types, e.g., type D and E. As people did not agree strongly on the five-type categorization, different audience description schemes were explored to obtain a more reliable description scheme. New description schemes were constructed by combining the types whose descriptions or labels might not be differentiated, e.g., type D and E. The coefficients

Table 6 Agreement coefficients between audience descriptions and labels using different audience description scheme

Audience description scheme	A, B, C, D, E	A, B, C, Inattentive (D and E)	Attentive (A, B, and C), Inattentive (D and E)
Kappa	0.48*	0.65*	0.73*

* $p < 0.001$

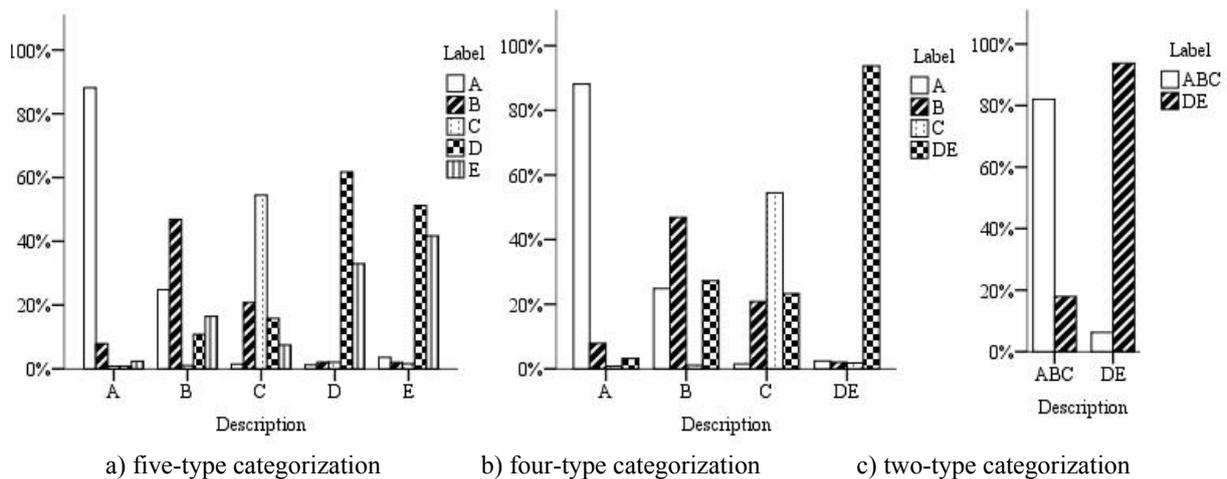


Fig. 5. An overview of description-label relationship using different audience categorization schemes. The label choice distributions are expressed as a percentage for each description (category), i.e., the percentages of labels for one situation description (grouped bars filled with different patterns) should add up to 1.

(Table 6) increased to an acceptable level of 0.73 when audience types A, B, and C were integrated into one category and D and E into another, creating an attentive and an inattentive audience group. Thus, people showed more agreement on the two-type categorization. This suggested that most noticeable feature in the audience behavior was whether or not an audience paid attention to the presenter.

Table 7 Number of participants ($n = 100$) who categorized the audience as Inattentive

Audience type	Sample 1		Sample 2	
	Description	Label	Description	Label
A	17	23	11	14
B	12	16	14	21
C	22	29	18	18
D	75	87	73	86
E	63*	80	68	82

* $p=0.12$, the p value of all other results without notation is $p<0.001$.

To investigate whether participants distinguished the different audience types, participants' perception was tested across the audience types, using the binominal description scheme, i.e., attentive and inattentive. The numbers of participants who categorized the types as inattentive were tested by binomial tests against a random proportion of 50%. The test was conducted respectively on the descriptions and the labels for each sample video. As shown in Table 7, audiences A, B, and C were mostly ($p < 0.001$) categorized as attentive while most ($p < 0.05$) people categorized D and E as inattentive.

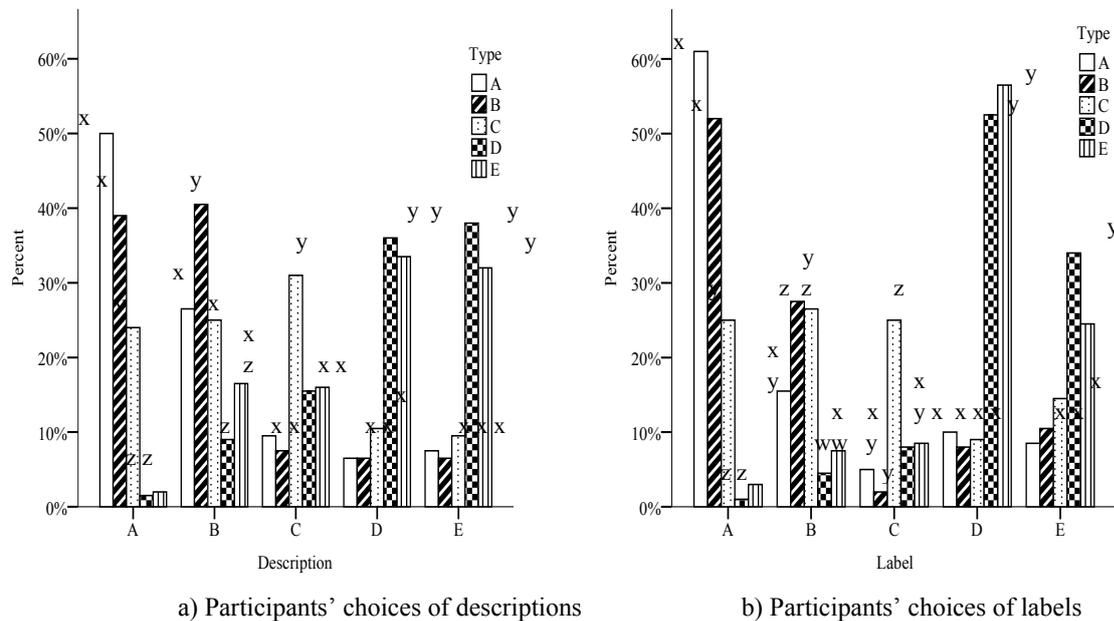


Fig. 6. An overview of participants' choices of descriptions and labels for different audience types. The description and label choice distributions are expressed as a percentage on each condition, i.e., the percentage of each description or label in one type (bars filled with the same pattern) should add up to 1. Each letter on top of a bar denotes a subset of audience type categories whose proportions for a certain description or label do not differ significantly ($p < 0.05$) from each other.

Besides the obvious difference between attentive audiences and inattentive audiences, the differences within each group were also investigated. Fig. 6 shows an overview of the choice distribution of descriptions and labels for different audience types. Friedman tests were conducted on each description and label choice to find out whether the choices varied across the audience types. The results (Table 8) show an overall significant ($p < 0.001$) difference in the choices of each description and each label among the audience types. Pairwise comparisons were further conducted using z-test on the proportions of each description or label between different types. The significance values of the z-tests had been adjusted using the Bonferroni method. The results are also shown in Fig. 6. Description A and label A were most preferred by the participants for type A, although not significantly more than description B and label B; description B was significantly ($p < 0.05$) more preferred for type B than other conditions while label B was preferred for both types B and C; both description and label C were selected by significantly ($p < 0.05$) more participants in type C; no significant difference was found between types D and E, with similar preference for descriptions and labels D and E. Therefore, both description and label results showed the differences within the attentive audiences (A, B, and C), but no difference was found within inattentive audiences (D and E). However, some labels did not show a one-to-one match for a certain type. For example, label B was found often chosen for both types B and C, and label D were more chosen than label E by the participants to describe both types D and E.

Table 8 Results of Friedman tests of description and label choices across the audience types, $\chi^2(4, N = 200)$

Description					Label				
A	B	C	D	E	A	B	C	D	E
197.33	60.57	48.51	115.83	120.31	275.23	62.63	71.87	239.21	56.57

Note: the p values of all statistics are less than 0.001.

In conclusion, the difference between attentive (A, B, and C) and inattentive (D and E) audiences was perceived to be significant. The descriptions and labels for types A, B, and C were partially validated: although people could not always perceive the difference between A, B, and C, there was a trend that the corresponding descriptions and labels were often the most preferred. No significant difference was found between types D and E.

6. Study IV: the behavior of the audience types

To gain some insight about people's perception of the audience types, the audience behavior was investigated for the different audience types. The study was conducted for each audience type on three aspects: (1) the specific postures, (2) the frequency of body movements, and (3) reactions towards disruptive events.

6.1 listening postures of the perceivable audience types

To examine the specific behavior for the perceivable audience types, the statistical model for the generation of listening postures was used. The median settings of five audience types obtained in Study II were used as the model input, and thus one posture category was obtained for each audience type.

Table 9 summarizes the audience behavior for each type. Concerning the head position, a trend of decreasing attention was observed in order of type A, B, C, and D. While the head for type A was always facing the front, two out of three head positions for type D were looking downwards. A trend of decreasing openness showed in the position of arms and hands from type A to C. Compared with type A, type B audience showed more closed gestures, e.g., clenched hands and folded arms, and the gestures in C are totally closed. Type C also differed from A and B in the torso position. The upright position suggested less relaxation in the torso. Apart from the head position, type D also distinguished well from A, B, and C in arm and leg position, e.g., the fidgeting hands and legs. The behaviors for type D and E were almost the same except the facial expression. However, as the audience's heads were mostly lowered, it might explain that a difference was hardly recognized in Study III.

Table 9 audience behavior in different settings

		A Interested and enthusiastic	B Critical and concerned	C Anxious and threatened	D Bored and impatient	E Indifferent and uninterested
Mood setting	Valence	H	M	L	L	M
	Arousal	H	M	H	L	L
	Dominance	M	M	M	M	M
						
Listening Posture	Number of postures	2	4	3	3	3
	Head	Upright	³ Upright; ¹ tilted position, facing the front	¹ Upright; ¹ lowered head ¹ tilted head	¹ Upright; ² lowered head	¹ Upright; ² lowered head
	Arms and hands	¹ hands open on desk; ¹ one hand touching or holding the other arm	¹ clenched hands resting on desk; ¹ hands open on desk; ¹ one hand touching the neck, with the other resting on the front torso; ¹ folded arms	² arms folded; ¹ hands on legs	² supporting the head; ¹ one or two hands tap on the desk continuously	² supporting the head; ¹ one or two hands on tap the desk continuously
	Torso	¹ Torso forward; ¹ Torso backward.	Torso backward	Torso upright	Torso forward	Torso forward
	Legs	Crossed or twisted legs.	Crossed or twisted legs.	Crossed or twisted legs.	² standard position – both feet flat on the floor; ¹ leg joggling or tapping on the floor	² standard position – both feet flat on the floor; ¹ leg joggling or tapping on the floor
Average probability of posture shifts		0.03	0.04	0.11	0.06	0.06
Reaction to an Event no longer than four seconds		No	React	React	React	React
Reaction to an Event no shorter than five seconds		React	React	React	React	React

Note: the superscript numbers indicate how many postures of the corresponding audience type comprise this position. If a position is not specified, all the postures of this audience type comprise this position.

6.2 Bodily movements

Another factor that affects the perception of audience behavior is the frequency of bodily movements. The bodily movements include posture shifts and consistently moving behaviors such as finger tapping. To study how often the behavior shifted from one posture to another, the posture transition matrix of the audience behavior model was inspected. The transition matrix presented the probabilities for each posture to transition to other postures in the

successive time unit, i.e., two seconds in the model. Thus, the probabilities for posture shifts were used as one measure of bodily movement. To judge how the frequencies differed across the audience types, the average of the probabilities within one posture category for each audience type was calculated as the measure (also shown in Table 9). The results present an increase in the probability of posture shifts in the order of audience type A, B, and C. Although type D audience shifted their postures less often than type C, it actually exhibits much more bodily movements, because two out of three postures involve consistently moving behaviors such as finger tapping. As type D and E employed the same posture category, the movement probability was also the same. This suggested that an inattentive audience might exhibit more bodily movements than an attentive audience.

6.3 Event reaction

Besides listening behavior, the different audience settings may also affect the audience member reaction to disturbing events, specifically, whether or not to respond to such an event, such as a door slam. To study the reaction, the audience settings were used as the input of the reaction decision function (Kang et al., 2013). As the reaction was also related to the event duration, the event duration was set 0.5, 1, 2, 3, 4, and 5 seconds respectively. The results (shown in Table 9) show that when the event duration was short, i.e., no longer than four seconds, only type A audience would not respond, while all other types would respond. However, if an event was long enough to be distracting, the most attentive audience would also respond.

7. Discussion and Conclusions

In conclusion, people were able to perceive changes in some of mood, personality, and attitude parameters by observing a virtual audience's behavior. Using the perceivable parameters, several audience scenarios were constructed by a group of individuals who acted as designers of virtual audiences. Audience parameter settings of individual audience scenarios showed extensive consistency across these designers. Furthermore, the list of 21 audience scenarios could be clustered into five underlying generic audience behavior styles. This led to the creation of five audience styles using the median settings of each cluster group. The perception validation study of these five styles showed a dominating characteristic of an audience that people perceived was whether or not the audience was attentive or inattentive. Although weaker, the findings also suggested that people could distinguish between interested-enthusiastic audience, critical-concerned audience and anxious-threatened audience. Finally, the findings of study IV gave an overview of the audience behaviors that made up these five audience styles. We anticipate that future developers can use these to create different recognizable audience styles.

As suggested in the previous study (Kang et al., 2013), facial expressions were added to improve the virtual audience's expressiveness. This point was supported by the results of Study I and Study II. However, according to the rationale provided by participants in the open question in Study III, it seemed that many participants did not use facial expressions as a clue to the judgment of audience styles. For example, only 18 out of 100 participants mentioned facial expression for videos of interested-enthusiastic (type A) audience. This might be caused by several reasons. First, the facial expressions in Study III were static in each condition. Without any change or comparison, people might not differentiate between some expressions, e.g., moderate pleasure and a neutral mood. Second, there was no control of the display devices the participants used to watch animations in the online survey. If the screen was small, participants might not have seen the facial expressions clearly as the scene included 11 characters in total. Nevertheless, Study III also showed that without any hardware constraints, people could recognize different audience styles. Besides facial expressions, another explanation of the lack of expressiveness in some situations is the underlying audience model used in the studies. As it was built based on observations of only 16 people in a university, the output behavior may lack variations for some audience types such as D and E.

This study can be extended in many directions. First, behavior of the current audience corpus was based on observations of normal conditions without extreme moods or attitudes. The behavioral model can be extended to show more extreme conditions by observing audiences in more diverse situations other than the classroom setting in the current corpus, e.g., business meeting and theatre. It is also worth exploring whether audiences' social-economic

backgrounds or cultures influence their behavioral styles and whether people from more similar social-economic backgrounds or cultures would show more agreement about their perception of these audience styles. Second, to make a clear distinction between different audience styles, the parameter settings were always similar for all virtual audience members in the presented studies. However, such a homogeneous audience would hardly exist in real life. Studying a more heterogeneous audience would therefore provide more insights into behaviors and people's perception of more complex real life audiences. Third, in the studies there was no interaction between audience and the presenter. It would therefore be interesting to study a responsive virtual audience that would react according to the speaker's behavior. In this way, a presenter may perceive a stronger connection with the virtual audience, hence higher social presence (Biocca & Harms, 2002). Besides the perception of a virtual audience from bystanders' view as conducted in this paper, speaker's perception and responses could also be investigated, involving factors such as speech content, emotions, and the speaker's confidence.

When virtual audiences take the place of real audiences for various purposes such as psychotherapy and performance rehearsal, it is important that virtual audiences elicit similar responses in the users to those elicited by real audiences to ensure the effectiveness. For this, immersion, place illusion, and plausibility illusion are the three key concepts to understand (Slater, 2009). Whereas immersion is a description of the characteristics of the system, e.g. the image quality of virtual audience, presence, i.e. place illusion, is related to the feeling of 'being there', and plausibility illusion refers to the illusion that the depicted scenario is actually occurring. A recent meta-analysis (Ling, Nefs, Morina, Heynderickx, & Brinkman, 2014) on the relationship between presence and the intended provoked anxiety within virtual environments developed for psychotherapy of anxiety disorders, however, found no correlation between anxiety and presence experienced in virtual environments for treatment of social anxiety such as virtual audiences. Ling et al (2014) pointed out that the presence measured in the studies mainly considered space illusion but not plausibility illusion, which might be a key issue when it came to social anxiety. In a similar manner, Poeschl and Doering (2013) also stressed the need to understand people's experience of realism when exposure to scenario that involved a virtual audience. Future work therefore should focus on measuring the plausibility illusion when studying virtual audience, i.e. the illusion that the social interaction with the virtual audience is actually happening. For social situations such as public speaking, this illusion specifically relates to social presence which refers to users' perception on the virtual social characters and experience of their relationship with the virtual characters. Thus, social presence with virtual audiences could also be an important factor which affects users' responses to virtual audiences.

In conclusion, this paper explored audience simulation parameters, their settings and consequent audience styles, and validated them through a series of perception studies. This contribution is important because virtual audiences often function as key stimulus material. Validation is also vital as it provides the foundation for drawing any valid conclusions later on about people's behavior, emotions, and attitudes when they are exposed to these virtual audiences. The work also presents a more practical contribution by providing developers with guidelines for designing the behavior of virtual audiences. The potential existence of at least five underlying audience styles among the 21 public scenarios suggests that the five styles could represent a large variety of audiences which would occur in various public speaking scenarios. Thus, by implementing only five audience styles, designers would be able to construct many more different social settings with an audience, and users would have opportunities to experience more variations of social settings. As the parameter settings of the five audience styles also show consistency in the virtual audience's moods, attitudes, and personalities, designers should also consider the expressiveness of a virtual audience as a key factor to construct different audience styles successfully. Besides, as an audience's attentiveness is suggested as a dominating perceivable characteristic, it is an important characteristic to be mentioned and considered when describing or designing an audience. Additionally, the specific postures and behavioral patterns found in the five audience styles may help designers to develop virtual audiences with noticeable and recognizable behavioral styles. The findings can also be generalized to the design of individual virtual characters acting as listeners. Specifically, to design expressive virtual listeners, their behavior should variate in the following aspects: head and gaze direction, facial expression, frequency of bodily movements, reaction to disturbing events, and postural features

such as openness, relaxation, and fidgets. The findings of the last study give designers directions on how to modulate these behaviors to create listening individuals as well as complete virtual audiences.

Reference

- Anderson, P. L., Price, M., Edwards, S. M., Obasaju, M. a, Schmertz, S. K., Zimand, E., & Calamaras, M. R. (2013). Virtual reality exposure therapy for social anxiety disorder: a randomized controlled trial. *Journal of consulting and clinical psychology, 81*(5), 751–60. doi:10.1037/a0033559
- Aymerich-Franch, L., Kizilcec, R. F., & Bailenson, J. N. (2014). The Relationship between virtual self similarity and social anxiety. *Frontiers in Human Neuroscience, 8*, 944. doi:10.3389/fnhum.2014.00944
- Bautista, N. U., & Boone, W. J. (2015). Exploring the impact of TeachME™ Lab virtual classroom teaching simulation on early childhood education majors' self-efficacy beliefs. *Journal of Science Teacher Education, 26*(3), 237–262. doi:10.1007/s10972-014-9418-8
- Bevacqua, E., Sevin, E., Hyniewska, S. J., & Pelachaud, C. (2012). A listener model: Introducing personality traits. *Journal on Multimodal User Interfaces, 6*(1-2), 27–38. doi:10.1007/s12193-012-0094-8
- Bissonnette, J., Dubé, F., Provencher, M. D., & Moreno Sala, M. T. (2015). Virtual reality exposure training for musicians: Its effect on performance anxiety and quality. *Medical problems of performing artists, 30*(3), 169–77. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/26395619>
- Biocca, F., & Harms, C. (2002). Defining and measuring social presence: Contribution to the networked minds theory and measure. In F. R. Gouveia & F. Biocca (Eds.), *Proceedings of the 5th International Workshop on Presence* (Vol. 2002, pp. 1–36). Citeseer.
- Broekens, J., Qu, C., & Brinkman, W.-P. (2012). *Dynamic facial expression of emotion made easy* (Technical report). Retrieved from <http://www.joostbroekens.com/>
- Chollet, M., Ochs, M., & Pelachaud, C. (2014). From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Proceeding of the 14th International Conference on Intelligent Virtual Agents* (pp. 120–133). Springer International Publishing.
- Chollet, M., Sratou, G., & Shapiro, A. (2014). An interactive virtual audience platform for public speaking training. *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* (pp. 1657–1658).
- Engeldrum, P. (2000). *Psychometric scaling: A toolkit for imaging systems development*. Imcotek Press.
- Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style*. Johns Hopkins University Press.
- Hartanto, D., Brinkman, W.-P., Kampmann, I. L., Morina, N., Emmelkamp, P. G. M., & Neerinx, M. A. (2015). Design and implementation of home-based virtual reality exposure therapy system with a virtual ecoach. In W.-P. Brinkman, J. Broekens, & D. Heylen (Eds.), *Proceeding of the 15th International Conference on Intelligent Virtual Agents* (Vol. 9238, pp. 287–291). Springer International Publishing. doi:10.1007/978-3-319-21996-7_31
- Heimberg, R. G., & Becker, R. E. (2002). *Cognitive-behavioral group therapy for social phobia: basic mechanisms and clinical strategies*. Guilford Publications.

- Hofmann, S. F., & Otto, M. W. (2008). *Cognitive behavioral therapy for social anxiety disorder: Evidence-based and disorder-specific treatment techniques* (1st ed.). Routledge. doi:10.1007/SpringerReference_179876
- Hu, C., Walker, M. A., Neff, M., & Tree, J. E. F. (2015). Storytelling agents with personality and adaptivity. In W.-P. Brinkman, J. Broekens, & D. Heylen (Eds.), *Proceeding of the 15th International Conference on Intelligent Virtual Agents* (Vol. 9238, pp. 181–193). Springer International Publishing. doi:10.1007/978-3-319-21996-7
- Kang, N. (2013). Posture corpus for the behavior generator [Dataset]. doi:10.4121/uuid:d613cc9c-c10b-4c50-be50-ba8ef7885dc5
- Kang, N., Brinkman, W.-P., Van Riemsdijk, B., & Neerinx, M. A. (2013). An expressive virtual audience with flexible behavioral styles. *IEEE Transactions on Affective Computing*, 4(4), 326–340. doi:10.1109/TAFFC.2013.2297104
- Kelly, O., Matheson, K., Martinez, A., Merali, Z., & Anisman, H. (2007). Psychosocial stress evoked by a virtual audience: Relation to neuroendocrine activity. *CyberPsychology & Behavior*, 10(5), 655–62. doi:10.1089/cpb.2007.9973
- Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The “Trier Social Stress Test”- A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2), 76–81.
- Ling, Y., Nefs, H. T., Morina, N., Heynderickx, I., & Brinkman, W.-P. (2014). A meta-analysis on the relationship between self-reported presence and anxiety in virtual reality exposure therapy for anxiety disorders. *PloS one*, 9(5), e96144. doi:10.1371/journal.pone.0096144
- Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4), 261–292. doi:10.1007/BF02686918
- Morina, N., Brinkman, W.-P., Hartanto, D., Kampmann, I. L., & Emmelkamp, P. M. G. (2015). Social interactions in virtual reality exposure therapy: A proof-of-concept pilot study. *Technology and health care : official journal of the European Society for Engineering and Medicine*, 23(5), 581–9. doi:10.3233/THC-151014
- Pertaub, D. P., Slater, M., & Barker, C. (2002). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators & Virtual Environments*, 11(1), 68–78. doi:10.1162/105474602317343668
- Pervin, L. A., Cervone, D., & John, O. P. (2005). *Personality: Theory and Research* (9th ed.). John Wiley & Sons.
- Poeschl, S., & Doering, N. (2012). Designing virtual audiences for fear of public speaking training - An observation study on realistic nonverbal behavior. In B. K. Wiederhold & G. Riva (Eds.), *Annual Review of Cybertherapy and Telemedicine* (pp. 218–222). Interactive Media Institute and IOS Press.
- Poeschl, S., & Doering, N. (2013). The German VR Simulation Realism Scale--psychometric construction for virtual reality applications with virtual humans. *Studies in health technology and informatics*, 191, 33–7.
- Powers, M., & Emmelkamp, P. (2008). Virtual reality exposure therapy for anxiety disorders: A meta-analysis. *Journal of Anxiety Disorders*, 22(3), 561–569.
- Rajae-Joordens, R., & Engel, J. (2005). Paired comparisons in visual perception studies using small sample sizes. *Displays*, 26(1), 1–7. doi:10.1016/j.displa.2004.09.003

- Slater, M., Pertaub, D.-P., Barker, C., & Clark, D. M. (2006). An experimental study on fear of public speaking using a virtual environment. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, 9(5), 627–33. doi:10.1089/cpb.2006.9.627
- Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1535), 3549–57. doi:10.1098/rstb.2009.0138
- Taylor, S. E., Seeman, T. E., Eisenberger, N. I., Kozanian, T. A, Moore, A. N., & Moons, W. G. (2010). Effects of a supportive or an unsupportive audience on biological and psychological responses to stress. *Journal of Personality and Social Psychology*, 98(1), 47–56. doi:10.1037/a0016563
- Thalmann, D., & Musse, S. R. (2013). *Crowd Simulation*. London: Springer London. doi:10.1007/978-1-4471-4450-2
- Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, 34, 273–286.
- Tudor, A.-D., Poeschl, S., & Doering, N. (2013). What do audiences do when they sit and listen? In B. K. Wiederhold & G. Riva (Eds.), *Annual Review of Cybertherapy and Telemedicine* (pp. 120–124). IOS press.
- Wallergard, M., Jonsson, P., Osterberg, K., Johansson, G., & Karlson, B. (2011). A virtual reality version of the Trier Social Stress Test: A pilot study. *Presence: Teleoperators and Virtual Environments*, 20(4), 325–336.
- Wang, Z., Lee, J., & Marsella, S. (2011). Towards more comprehensive listening behavior: Beyond the bobble head. In H. H. Vilhjálmsson, S. Kopp, S. Marsella, & K. R. Thórisson (Eds.), *Proceedings of the 11th international conference on Intelligent Virtual Agents* (Vol. 6895, pp. 216–227). Berlin, Heidelberg: Springer-Verlag.
- Xu, J., Broekens, J., & Hindriks, K. (2013). Bodily mood expression: Recognize moods from functional behaviors of humanoid robots. In G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *proceedings of 5th international conference on social robots (ICSR)* (pp. 511–520). Bristol, UK: Springer International Publishing.
- Xu, Y., Pelachaud, C., & Marsella, S. (2014). Compound gesture generation: A Model based on ideational units. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Proceeding of the 14th International Conference on Intelligent Virtual Agents* (pp. 477–491). Springer International Publishing.
- Zanbaka, C., Ulinski, A., Goolkasian, P., & Hodges, L. F. (2007). Social responses to virtual humans: implications for future interface design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1561–1570). ACM.

Appendix A

The principal component analysis was conducted with varimax orthogonal rotation. Five factors were extracted with eigenvalues over Kaiser's criterion of 1 and in combination explained 74% of the variance. The factor analysis describes how the parameters correlate with the extracted factors. Hence, the parameters with similar factor loadings indicated high correlations between each other, thereby being grouped together (Table A.1). As the factors were independent of each other, the parameter groups were independent of each other if they correlated with different single factors. Thus, there were three independent parameter groups (IG1, IG2, and IG3) and two independent single-parameter groups (IP1 and IP2). Each factor could be interpreted as a characteristic presented by its correlated independent parameters. For example, factor 2 could be interpreted as Patience or Eagerness for information. For the three parameters (DP1, DP2, and DP3) which were correlated with multiple factors, the parameters could also be

explained by the correlated factors. For example, the value of Approval (DP2) correlates positively with the values of Eagerness for information (factor 2) and Dominance (factor 3) and negatively with Criticism (factor 4).

Table A.1 Factor loadings of audience parameters and grouping result

Parameters	Loadings on factors					Grouping result
	1	2	3	4	5	
Extraversion	0.907	0.036	0.044	0.023	0.093	
Agreeableness	0.805	-0.068	-0.002	-0.140	-0.080	IG1
Conscientiousness	0.670	0.137	0.277	-0.024	-0.234	
Openness	0.689	0.051	0.229	0.408	0.307	DP1
Impatience	-0.012	-0.848	-0.091	-0.084	0.027	
Eagerness for info	0.090	0.823	-0.185	0.133	-0.126	IG2
Interest	0.004	0.791	0.276	-0.023	-0.115	
Approval	-0.003	0.517	0.410	-0.495	0.231	DP2
Dominance	0.176	-0.043	0.856	0.043	-0.054	IG3
Valence	0.010	0.376	0.672	-0.207	0.271	
Neuroticism	-0.420	-0.070	-0.577	-0.341	0.225	DP3
Criticism	-0.041	0.153	0.032	0.856	0.018	IP1
Arousal	-0.030	-0.167	0.000	-0.010	0.911	IP2

Note: the loadings larger than 0.4 are in bold type.

Appendix B

The basic method of paired comparisons consists of sequentially presenting pairs of stimuli to an observer and asking the observer which one of the pair has the greatest amount of a certain attribute. Supposing there are n stimuli to compare in total, each observer will have a $n \times n$ matrix of comparison results. If the observer selects stimulus j over i , as having more of the attribute in question, we put a 1 in the j^{th} column and the i^{th} row of a matrix. Using all the matrices of the J observers, a frequency matrix, F , was accumulated. In this matrix, each element is the number of times the stimulus in the j^{th} column was chosen over the stimulus in the i^{th} column. The next step is to form the proportion matrix, P , by dividing each element of F by the number of observers, J . That is, each element of P , i.e., $p_{j>i}$, represents the proportion of observers who select stimulus j over i .

According to Case V of Thurstone model, the scale value difference of two stimuli j and i (i.e., $S_j - S_i$) can be expressed as the z-score corresponding to the preference frequency (or proportion) of stimulus j over i , $p_{j>i}$. The formula is shown below:

$$S_j - S_i = Z_{j>i} = F^{-1}(p_{j>i}),$$

where F^{-1} is the inverse of the standard cumulative normal distribution function.

By transforming each element in the P matrix into a corresponding z-score, a matrix, S , of scale value differences, is then obtained, shown as follows:

$$S = \begin{bmatrix} S_1 - S_1 & S_2 - S_1 & \cdots & S_n - S_1 \\ S_1 - S_2 & S_2 - S_2 & \cdots & S_n - S_2 \\ S_1 - S_3 & S_2 - S_3 & \cdots & S_n - S_3 \\ \vdots & \vdots & \cdots & \vdots \\ S_1 - S_n & S_2 - S_n & \cdots & S_n - S_n \end{bmatrix}$$

The scale values of each stimulus can be determined from the column sums of the S matrix. Taking the first column for example, by dividing the column sum by the number of stimuli, we have $\frac{1}{n} \sum_{i=1}^n (S_1 - S_i) = S_1 - \bar{S}$. As the average of all the scale values can be set zero, i.e., $\bar{S} = 0$, the column sums give the scale values directly, i.e., $S_1 - 0 = S_1$.

Appendix C: Descriptions of audience scenarios

Note: Only full scenario descriptions were provided to the participants; the short descriptions are used for convenience when the full descriptions are referred to in this paper.

No.	Short description	Full scenario description
1	Promising business proposal	A person wants to start his own company and needs a sizable amount of investment money for this. He has an opportunity to introduce the investment proposal within 10 minutes to a number of business people, as they will consider whether or not they might invest in this new business opportunity. While listening, the investors find the proposal very promising.
2	Potential business proposal	A person wants to start his own company and needs a sizable amount of investment money for this. He has an opportunity to introduce the investment proposal within 10 minutes to a number of business people, as they will consider whether or not they might invest in this new business opportunity. While listening, the investors find the proposal has some potential but still has some concerns about a number of issues that would require additional work for the person to work out in more detail.
3	Impractical business proposal	A person wants to start his own company and needs a sizable amount of investment money for this. He has an opportunity to introduce the investment proposal within 10 minutes to a number of business people, as they will consider whether or not they might invest in this new business opportunity. While listening, the investors find the proposal definitely impractical.
4	Best man's talk	The best man is talking about an interesting story about the new couple at a wedding party. The people in the party are mostly the new couples' family members and friends. They are friendly and enjoy very much the stories.
5	Friday afternoon lecture	On a Friday afternoon, a teacher is talking about an interesting example in a course. As the content will not appear in the coming exam, the students are more eager to leave.
6	Tuesday morning lecture about exam	On a Tuesday morning, 10 am, a teacher is talking about an interesting example in a course. The content will appear in the coming exam.
7	Attractive Tuesday morning lecture	On a Tuesday morning, 10 am, a teacher is talking about an interesting example in a course. The content attracts the students.
8	Unrelated Monday morning meeting	An employee is presenting his/her work during a meeting with a dozen of colleagues on a Monday morning. The attendees are fellow employees but work on non-related projects. They are indifferent to what the employee is presenting.
9	Related Monday morning meeting	An employee is presenting his/her work during a meeting with a dozen of colleagues on a Monday morning. The attendees are fellow employees working on related projects, so they like to learn how this might affect or benefit their project.
10	Funny show	A beginning stand-up comedian is doing a prepared show for a crowd. The audience is enjoying the jokes and laughing.
11	Not funny show	A beginning stand-up comedian is doing a prepared show for a crowd. The audience is indifferent and finds it not funny at all, and some people are eager to leave.
12	Funeral eulogy	A person is delivering a eulogy at a funeral. The attendees are the family members and best friends of the deceased.

13	Souvenir introduction	At the end of a one-day guided tour, a salesman is introducing a souvenir to the tourists. There is nothing special with the souvenir, and the tourists hope to finish as soon as possible.
14	Positive corporate report	A director of a small company presents a corporate report to all the 11 employees. The report shows improved performance on key targets and they will receive a big annual bonus.
15	Repeated rule announcement	During a weekly school assembly for high school students, the administrator talks about the rules the students should obey, which he or she repeated every week.
16	Announcement of stricter rules	During a weekly school assembly for high school students, the administrator is talking about the new rules the students should obey. The students find the rules much stricter than before.
17	Budget cut	A company manager announces to the work team that a few employees will be made redundant because their budget has been cut. The employees are nervous.
18	Innovative design	At a booth of an exhibition, an exhibitor is introducing a new product. People follow the explanation and find the design innovative.
19	Looking around in an exhibition	At a booth of an exhibition, an exhibitor is introducing a new product. People may only be interested to look around or only to get the freebies.
20	A qualified interviewee	A person is giving a presentation as part of a job interview. The employers find the person sufficiently qualified.
21	A not very satisfactory interviewee	A person is giving a presentation as part of a job interview. The employers find the person may be qualified, but they are not very satisfied with him or her at some points.
22	Training plan	A presentation to fellow sport members at a meeting of the local sport association, about a new idea of this year's training plan.
23	Hobby talk	A student gives a half-hour talk to some fellow students. The student is talking about his hobby which some of the other students also like, but not all.

Appendix D

Suppose X_i and Y_i ($i = 1, 2, \dots, 6$) are the parameters for scenario settings X and Y , thereby $X = (X_1, X_2, \dots, X_6)$ and $Y = (Y_1, Y_2, \dots, Y_6)$. The observed distance on parameter i ($i = 1, 2, \dots, 6$) is noted as d_{O_i} , and the observed distance between setting X and Y is noted as d_O . The expected distance of d_{O_i} and d_O are respectively noted as d_{E_i} and d_E .

The observed distance for one parameter, d_{O_i} , is calculated as follows:

$$d_{O_i} = |X_i - Y_i|, \quad i = 1, 2, \dots, 6.$$

The observed distance between two settings, d_O , is the Euclidean distance between X and Y , and can be calculated from the distances for all the six dimensions, i.e., the six parameters listed in Table 2:

$$d_O = |X - Y| = \sqrt{\sum_{i=1}^6 (d_{O_i})^2}.$$

To calculate the expected distance between two settings, the expected distances on the six parameters were first calculated. The expected distance for parameter i , d_{E_i} , ($i = 1, 2, \dots, 6$) is calculated according to the distribution of the possible distances for one parameter. The possible values of d_{O_i} are shown in the Table C.1.

Table C.1 the possible values of the distance for one parameter, $d_{O_i} = |X_i - Y_i|$, $i = 1, 2, \dots, 6$

		X_i		
		L	M	H
Y_i	L	0	5	10
	M	5	0	5
	H	10	5	0

* The values for L, M, and H were respectively 0, 5, and 10.

As the possible values of a parameter show an equal distribution, i.e., a probability of 1/3 for L, M, and H respectively, the probability for any possible combination of X_i and Y_i is as follows:

$$P(X_i = x, Y_i = y) = \frac{1}{3} * \frac{1}{3} = \frac{1}{9}, x, y \in \{L, M, H\}.$$

That is, the probability for any possible distance listed in Table 3 is 1/9. Thus, the expected distance for one parameter d_{E_i} ($i = 1, 2, \dots, 6$) is calculated subsequently:

$$d_{E_i} = E[|X_i - Y_i|] = \sum_{x,y \in \{L,M,H\}} |x - y| \cdot P(X_i = x, Y_i = y) = (0 + 5 + 10 + 5 + 0 + 5 + 10 + 5 + 0) * \frac{1}{9} = \frac{40}{9}.$$

Hence the expected distance between two settings, d_E , is calculated from the distances for all the six dimensions, i.e., the six parameters listed in Table 2:

$$d_E = E[|X - Y|] = \sqrt{\sum_{i=1}^6 (d_{E_i})^2} = 10.88.$$